Sparse Subspace Clustering

Based on *Sparse Subspace Clustering: Algorithm, Theory, and Applications* by Elhamifar and Vidal (2013)

Alex Gutierrez

CSCI 8314

March 2, 2017

Outline

Outline

1. **Motivation and Background**
2. **Set-up and Algorithm**
3. **Practical Extensions**
4. **Theoretical Guarantees**
5. **Real Experiments**

Slides Based on *Sparse Subspace Clustering: Algorithm, Theory, and Applications*
by Elhamifar and Vidal (2013, algorithm first appeared in 2009)

Table of Contents

Data in low-dimensional subspaces

High dimensional data is often well-approximated by low-dimensional subspaces. For example:

1. Feature trajectories of a rigidly moving object in a video
2. face images of a subject under varying illumination
3. a hand-written digit with different rotations, translations, and thicknesses

are all well-modeled by a low-dimensional subspace of the corresponding ambient space.

Thus when collecting data from multiple classes we would expect the data to lie in the union of multiple subspaces.

Subspace clustering

Subspace clustering (at a high level)

Given many (noisy) data points drawn from a union of subspaces, find the subspaces and the assignments of each point to a subspace.

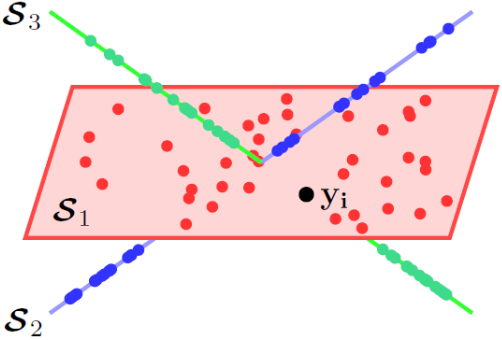


Table of Contents

Ingredients

1. **Ambient dimension:** $\mathbb{R}^d$
2. **Subspaces:** $\{S_\ell\}_{\ell=1}^n$, linear subspaces of $\mathbb{R}^d$
3. **Subspace dimensions:** $\{d_\ell\}_{\ell=1}^n$ the dimension of each subspace
4. **Data:** $\{y_i\}_{i=1}^N$, N data points lying in $\bigcup_{\ell=1}^n S_\ell$:

$$Y := [y_1, \dots, y_N] = [Y_1, \dots, Y_n]\Gamma,$$

where $Y_\ell \in \mathbb{R}^{D \times N_\ell}$ is the matrix containing all the points lying in S_ℓ and Γ is an (unknown) permutation matrix.

Ingredients

1. **Ambient dimension:** $\mathbb{R}^d$
2. **Subspaces:** $\{S_\ell\}_{\ell=1}^n$, linear subspaces of $\mathbb{R}^d$
3. **Subspace dimensions:** $\{d_\ell\}_{\ell=1}^n$ the dimension of each subspace
4. **Data:** $\{y_i\}_{i=1}^N$, N data points lying in $\bigcup_{\ell=1}^n S_\ell$:

$$Y := [y_1, \dots, y_N] = [Y_1, \dots, Y_n]\Gamma,$$

where $Y_\ell \in \mathbb{R}^{D \times N_\ell}$ is the matrix containing all the points lying in S_ℓ and Γ is an (unknown) permutation matrix.

Assume: Y_ℓ is rank d_ℓ and $N_\ell > d_\ell$.

Parsimonious Representation

Idea:

we should[1] be able to represent every point in a subspace as the linear combination of **just a few** other points in that same subspace.

Parsimonious Representation

Idea:

we should[1] be able to represent every point in a subspace as the linear combination of **just a few** other points in that same subspace.

[1] If we have enough data points in each subspace and the dimensions of each subspace is comparatively small

Parsimonious Representation

Idea:

we should[1] be able to represent every point in a subspace as the linear combination of **just a few** other points in that same subspace.

[1] If we have enough data points in each subspace and the dimensions of each subspace is comparatively small

Formally, given $y_i \in S_\ell$ we can write

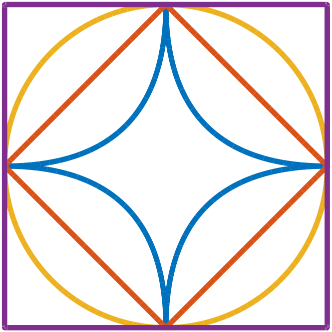
$$y_i = Yc_i, \qquad \text{where } c_{i,i} = 0$$

and $c_{i,j} \neq 0 \iff y_j \in S_\ell$.

Aside: how to promote sparsity

We would like to minimize the number of non-zero entries (the **"zero norm"**) in the solution to some linear system

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \quad & \|x\|_0 \\ \text{subject to} \quad & \|Ax - b\|_2 \leq \varepsilon \end{aligned}$$



Proposed Optimization Problem

This naturally motivates the optimization problem

$$\begin{aligned} \underset{c_i}{\text{minimize}} \quad & \|c_i\|_q \\ \text{subject to} \quad & y_i = Y c_i, \qquad c_{i,i} = 0 \\ & i = 1, \ldots, N. \end{aligned}$$

Proposed Optimization Problem

This naturally motivates the optimization problem

$$\begin{aligned} \underset{c_i}{\text{minimize}} \quad & \|c_i\|_q \\ \text{subject to} \quad & y_i = Y c_i, \qquad c_{i,i} = 0 \\ & i = 1, \ldots, N. \end{aligned}$$

(setting $q = 1$ and re-writing using matrix notation)

$$\begin{aligned} \underset{C}{\text{minimize}} \quad & \|C\|_{1,1} \\ \text{subject to} \quad & Y = YC, \qquad \text{diag}(C) = 0 \end{aligned}$$

Using C to cluster

To find the clusters:

1. Consider $|C|$ as the adjacency matrix of some graph
2. Symmetrize (make undirected): $W := |C| + |C|^T$
3. Perform **spectral clustering** on W
   1. Build the symmetric nomalized graph Laplacian
   $$L = I - D^{-1/2} W D^{1/2}.$$
   2. calculate the n bottom eigenvectors $U := [u_1, \ldots, u_n]$ of L
   3. Apply k-means to the normalized rows of U.

In the ideal case we will get n connected components

$$W = \begin{bmatrix} W_1 & \cdots 0 & \\ \vdots & \ddots & \vdots \\ 0 & \cdots & W_n \end{bmatrix} \Gamma.$$

Table of Contents

Noisy Data Model

We assumed before that we were given N noise-free data points. Let's extend this to a model

$$y_i = y_i^0 + e_i^0 + z_i^0,$$

which can handle:

- sparse outlying entries: $\left\|e_i^0\right\|_0 \leq k$
- noise: $\left\|z_i^0\right\|_2 \leq \zeta.$

Practical Optimization Problem (I)

We note that we can sparsely represent y_i^0 as before:

$$y_i^0 = \sum_{j \neq i} c_{i,j} y_j^0.$$

Following our nose, we write

$$y_i = \sum_{j \neq i} y_j + e_i + z_i$$
$$e_i := e_i^0 - \sum_{j \neq i} c_{i,j} e_j^0$$
$$z_i := z_i^0 - \sum_{j \neq i} c_{i,j} z_j^0.$$

Practical Optimization Problem (II)

It is clear now what the optimization program should be:

$$\begin{aligned} \underset{C}{\text{minimize}} \quad & \|C\|_{1,1} + \lambda_e \|E\|_{1,1} + \frac{\lambda_z}{2} \|Z\|_F^2 \\ \text{subject to} \quad & Y = YC + E + Z \quad \text{and} \quad \operatorname{diag}(C) = 0 \end{aligned}$$

Table of Contents

Definitions and Notation

Independent Subspaces

A collection of subspaces $\{S_\ell\}_{\ell=1}^n$ is said to be **independent** if $\dim\left(\oplus_{\ell=1}^n S_\ell\right) = \sum_{\ell=1}^n \dim S_\ell$

Definitions and Notation

Independent Subspaces

A collection of subspaces $\{S_\ell\}_{\ell=1}^n$ is said to be **independent** if $\dim\left(\oplus_{\ell=1}^n S_\ell\right) = \sum_{\ell=1}^n \dim S_\ell$

Disjoint Subspaces

We will call a collection of subspaces $\{S_\ell\}_{\ell=1}^n$ **disjoint** if the intersection of any two of them only contains the origin.

Definitions and Notation

Independent Subspaces

A collection of subspaces $\{S_\ell\}_{\ell=1}^n$ is said to be **independent** if $\dim\left(\oplus_{\ell=1}^n S_\ell\right) = \sum_{\ell=1}^n \dim S_\ell$

Disjoint Subspaces

We will call a collection of subspaces $\{S_\ell\}_{\ell=1}^n$ **disjoint** if the intersection of any two of them only contains the origin.

Let Y_ℓ denote the N_ℓ points corresponding to a subspace S_ℓ and let $Y_{-\ell}$ denote the rest of the data points.

Independent Subspaces Theory

Theorem 1

Assume that we have n independent subspaces of full rank: $\text{rank}(Y_\ell) = d_\ell$. Then for every S_ℓ and every non-zero y in S_ℓ the ℓ_q minimization problem:

$$\begin{bmatrix} c^* \\ c_-^* \end{bmatrix} . = \arg\min \left\| \begin{bmatrix} c \\ c_- \end{bmatrix} \right\|_q \quad \text{s.t.} \quad y = [Y_\ell \; Y_{-\ell}] \begin{bmatrix} c \\ c_- \end{bmatrix},$$

for $q < \infty$ recovers a subspace-sparse reprsentation. That is, $c^* \neq 0$ and $c_-^* = 0$.

Independent Subspaces Theory

Theorem 1

Assume that we have n independent subspaces of full rank: $\operatorname{rank}(Y_\ell) = d_\ell$. Then for every S_ℓ and every non-zero y in S_ℓ the ℓ_q minimization problem:

$$\begin{bmatrix} c^* \\ c^*_- \end{bmatrix} . = \arg\min \left\| \begin{bmatrix} c \\ c_- \end{bmatrix} \right\|_q \quad \text{s.t.} \quad y = [Y_\ell \; Y_{-\ell}] \begin{bmatrix} c \\ c_- \end{bmatrix},$$

for $q < \infty$ recovers a subspace-sparse reprsentation. That is, $c^* \neq 0$ and $c^*_- = 0$.

Note that the condition on independent subspaces is completely unreasonable.

Disjoint Subspaces Theory (I)

For a vector x in the intersection of S_ℓ with $\oplus_{j\neq\ell} S_j$ define

$$\begin{aligned} a_\ell &:= \arg\min \|a\|_1 \qquad \text{s.t.} \qquad x = Y_\ell A \\ a_{-\ell} &:= \arg\min \|a\|_1 \qquad \text{s.t.} \qquad x = Y_{-\ell} A \end{aligned}$$

Disjoint Subspaces Theory (I)

For a vector x in the intersection of S_ℓ with $\oplus_{j\neq\ell} S_j$ define

$$\begin{aligned} a_\ell &:= \arg\min \|a\|_1 \qquad \text{s.t.} \qquad x = Y_\ell A \\ a_{-\ell} &:= \arg\min \|a\|_1 \qquad \text{s.t.} \qquad x = Y_{-\ell} A \end{aligned}$$

We define the exact recover condition:

$$\forall x \in S_\ell \cap \left(\oplus_{j\neq\ell} S_j\right), \quad x \neq 0 \Rightarrow \|a_i\|_1 < \|a_{-i}\|_1 \,. \tag{ERC}$$

Disjoint Subspaces Theory (II)

$$\forall x \in S_\ell \cap \left(\oplus_{j \neq \ell} S_j \right), \quad x \neq 0 \Rightarrow \|a_i\|_1 < \|a_{-i}\|_1 . \tag{ERC}$$

Disjoint Subspaces Theory (II)

$$\forall x \in S_\ell \cap \left(\oplus_{j \neq \ell} S_j\right), \quad x \neq 0 \Rightarrow \|a_i\|_1 < \|a_{-i}\|_1 . \tag{ERC}$$

Theorem 2

Given n disjoint subspaces of full rank, for every S_ℓ and every non-zero y in S_ℓ the ℓ_q minimization problem:

$$\begin{bmatrix} c^* \\ c_-^* \end{bmatrix} . = \arg\min \left\| \begin{bmatrix} c \\ c_- \end{bmatrix} \right\|_1 \quad \text{s.t.} \quad y = [Y_i \; Y_{-i}] \begin{bmatrix} c \\ c_- \end{bmatrix},$$

recovers a subspace-sparse reprsentation if and only if (ERC) holds.

Table of Contents

Face Clustering Example

Given many different pictures of faces with varying illuminations, we seek to identify which faces belong to the same person.

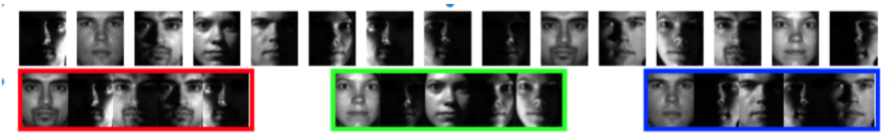


Fig. 2. Face clustering: given face images of multiple subjects (top), the goal is to find images that belong to the same subject (bottom).

(results take from Sparse Subspace Clustering)

Comparison to other methods

TABLE 1

Clustering error (%) of different algorithms on the Hopkins 155 dataset with the $2F$-dimensional data points.

Algorithms	LSA	SCC	LRR	LRR-H	LRSC	SSC
2 Motions						
Mean	4.23	2.89	4.10	2.13	3.69	**1.52** (2.07)
Median	0.56	**0.00**	0.22	**0.00**	0.29	**0.00** (0.00)
3 Motions						
Mean	7.02	8.25	9.89	**4.03**	7.69	4.40 (5.27)
Median	1.45	**0.24**	6.22	1.43	3.80	0.56 (0.40)
All						
Mean	4.86	4.10	5.41	2.56	4.59	**2.18** (2.79)
Median	0.89	**0.00**	0.53	**0.00**	0.60	**0.00** (0.00)

TABLE 2

Clustering error (%) of different algorithms on the Hopkins 155 dataset with the $4n$-dimensional data points obtained by applying PCA.

Algorithms	LSA	SCC	LRR	LRR-H	LRSC	SSC
2 Motions						
Mean	3.61	3.04	4.83	3.41	3.87	**1.83** (2.14)
Median	0.51	**0.00**	0.26	**0.00**	0.26	**0.00** (0.00)
3 Motions						
Mean	7.65	7.91	9.89	4.86	7.72	**4.40** (5.29)
Median	1.27	1.14	6.22	1.47	3.80	**0.56** (0.40)
All						
Mean	4.52	4.14	5.98	3.74	4.74	**2.41** (2.85)
Median	0.57	**0.00**	0.59	**0.00**	0.58	**0.00** (0.00)

(results take from Sparse Subspace Clustering)

Table of Contents

Practical Optimization Problem (III), affine subspaces

We can deal with affine spaces by including a single extra linear equality constraint:

$$\begin{aligned} \underset{C}{\text{minimize}} \quad & \|C\|_{1,1} + \lambda_e \|E\|_{1,1} + \frac{\lambda_z}{2} \|Z\|_F^2 \\ \text{subject to} \quad & Y = YC + E + Z, \qquad \text{diag}(C) = 0 \\ & \text{and} \quad 1^T C = 1^T, \end{aligned}$$

(simply adds on the equation for an affine plane)

Prior work on subspace clustering Part 0, iterative approaches

iterative approaches are largely based on generalizations of k-means

- k-subspaces (Tseng, 2000)
- median k-flats (Zhang, Szlam, and Lerman 2009)

Prior work on subspace clustering Part I, Algebraic

Algebraic approaches to subspace clustering involve either

- Factorization based approaches
  1. Costeira and Kanade, 1998
  2. Kanatani 2001
- Algebro-geometric approaches (called Generalized PCA)
  1. Vidal, Ma, and Sastry (2005)
  2. Ma, Yang, Derksen, and Fossum (2008)

Prior work on subspace clustering Part II: statistical approaches

These approaches generally assume a Gaussian distribution on the data inside each subspace and then use some basic estimation theory to find the spaces

- Mixtures of Probabilistic PCA (Tipping and Bishop, 1999)
- Multi-Stage Learning (EM based), (Gruber and Weiss, 2004)
- Random Sample Consensus (RANSAC, Fischler and Bolles, 1981)

Prior Work: Part III

Spectral clustering algorithms:

- Sparse Subspace Clustering (Elhamifar and Vidal 2009, Soltanolkotabi and Candes 2012)
- Low-rank recovery (LRR, Liu, Lin, and Yu 2010)
- Spectral Curvature Clustering of Chen and Lerman (2009)
- Local Subspace Affinity (LSA) by Yan and Pollefeys (2006)